A 130nm Generation High Density Etox™ Flash Memory Technology

Stephen N. Keeney

Intel Corporation
RN3-01, 2200 Mission College Blvd, Santa Clara, CA 95052
E-mail: stephen.n.keeney@intel.com

Abstract

A 130nm-generation flash memory technology has been developed, optimized for small cell size, high performance low voltage operation and multi-level-cell and embedded logic capability. Memory cell scaling utilizes the architecture features from the 180nm technology [1] along with channel erase, advanced 130nm lithography, dielectric scaling, junction scaling, dual trench and dual spacer technology. 32Mbit flash memories with a 0.16um^2 cell size have been built on this technology showing good yield, performance and reliability.

Introduction

The wireless communication market is driving the need for low cost high density and high performance non-volatile memory that can be embedded with a standard logic process. The goal for this generation of technology is to deliver half the die size of the previous generation at the same density using a logic compatible process. This paper will describe the flash technology that has been developed and will show some of the key results. Two key areas will be addressed which are the flash cell scaling and the periphery transistor scaling which constitute the majority of the die area.

Scaling Challenges

There are several key challenges to scaling a flash technology. For the flash cell itself, the source space, the gate length and the drain space determine the height of the cell. Reduction of the source space needs to be achieved without a significant increase in the source rail resistance, reduction of the gate length must occur while maintaining good punch through characteristics and scaling of the drain space must be achieved while maintaining good gap fill between the poly lines along with sufficient contact to gate margin. The scaling of the width of the device is limited by the physical to electrical difference and the isolation of neighboring cells. Scaling the periphery transistors can be achieved by reducing the maximum voltages that need to be supported along with junction engineering and more advanced lithography and etch capabilities. This paper will address each of these challenges resulting in a flash cell size of 0.16um^2.

Flash Cell Scaling

The flash cell scaling is enabled by three key modifications:

1) A dual trench scheme with a shallow trench in the flash array and a deep trench in the periphery (Fig.1)
2) A dual spacer scheme using a narrow spacer in the array and a wide spacer in the periphery (Fig.2)
3) Adoption of a channel erase scheme enabled by a triple well process.

Cell Height Scaling

One of the key challenges to source space scaling is managing the source rail resistance. The source implant dose and energy are carefully chosen to manage this resistance. However, moving to a dual trench scheme with the shallow trench in the array allows significant scaling of the source junction while maintaining a low resistance without compromising the periphery isolation capability. Along with the adoption of channel erase the source junction underlap of the gate is significantly reduced by junction scaling since this junction no longer needs to support high electric fields seen in the traditional negative gate erase scheme. This results in a reduction in the required physical gate length from the previous generation. The gate length can also be scaled with the reduction in the tunnel oxide thickness to 9nm, which provides better gate control of the channel. The spacer thickness is determined by the high voltage requirements of the periphery transistors. Choosing a dual spacer scheme allows a thin spacer to be used in the array to allow the flash drain spacer to be scaled without any new gap fill concerns while maintaining the high voltage

capability of the periphery transistors. These modifications are used in combination with advanced 130nm lithography for the gate patterning along with the contact definition to create a cell height of 495nm.

Cell Width Scaling

The dual trench scheme also allows the width of the flash cell isolation to be scaled. The shallow trench in the array enables the previous generation trench gap fill oxide to be used since the trench fill aspect ratio is not increased even with the reduction in isolation width. The self aligned poly (SAP) process along with the unlanded contact process from the previous generation are also required to prevent the floating gate or contact alignment being a concern (1). The trench corner rounding is still required to provide good oxide reliability but the oxidations used are scaled to reduce the amount of silicon loss which would result in a reduction in the active width of the flash cell. Advanced 130nm lithography is required for the Metal 1 patterning to fit into the pitch of the bitlines along with the small contacts that this delivers. The net result is a cell width of 330nm.

Transistor Optimization

To reduce cost the periphery transistors must also be scaled since they constitute a significant portion of the die area and accurate gate patterning is required to support an embedded logic capability. The introduction of channel erase reduces the maximum voltage the periphery needs to support and the introduction of more advanced lithography and etch gives better gate patterning capability. This allows the channel length and gate oxides to be scaled which done in conjunction with traditional junction scaling leads to a significant reduction in the gate length while maintain good transistor characteristics. For the embedded logic process this leads to a gate length of 100nm. The reduction in the maximum voltage the periphery needs to support along with the dual trench scheme allows the isolation width to be scaled as well since a deep trench can be maintained for logic devices independent of the shallow trench used in the flash array. These changes combined with the advanced 130nm lithography tools, Cobalt salicide and complimentary gates consistent with Intel's 130nm logic process (2) delivers the required transistor performance and area savings.

Results

Figures 3 and 4 show very fast program and erase characteristics achieved as a result of the tunnel oxide scaling, gate length scaling and junction optimization along with the triple well implementation in the array for channel erase. Fig.5 and Fig.6 show very good oxide quality demonstrating good trench corner integrity of the tunnel oxide as well as the high voltage periphery oxide even after scaling of the trench corner oxidations. Fig. 7 shows the distribution of the minimum bit in a 512k block over many blocks tested showing good retention and no anomalous charge loss tails even after 10k cycles and 500 hours of bake. Figure 8 shows the transistor characteristics versus back bias at minimum channel length showing no anomalous kink in the IV characteristics associated with poor trench corner rounding or the SAP process used here. This technology has been designed to support Multi Level Cell operation which will result in an effective cell size of $0.08um^2$ using 2 bits per cell.

Conclusion

An advanced 130nm generation flash memory process with a $0.16um^2$ flash cell area has been demonstrated to have good performance and reliability. Furthermore, the periphery transistors have also been optimized to meet the product performance requirements.

Acknowledgements

The author would like to thank CTM (California Technology and Manufacturing) group as well as FPG (Flash Products Group) for their contribution to this development program.

References

[1] A. Fazio, "A High Density High Performance 180nm Generation Etox™ Flash Memory Technology", IEEE IEDM Tech. Dig., 1999

[2] S. Tyagi et al. "A 130nm Generation Logic Technology Featuring 70nm Transistors, Dual Vt Transistors and 6 Layers of Cu Interconnects", IEEE IEDM Tech Deg. pp567-570, 2000

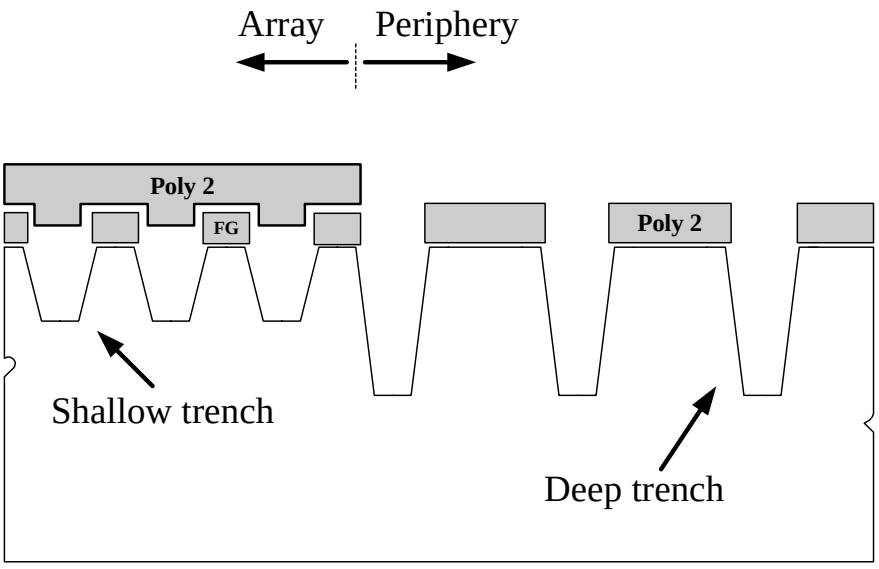

Fig. 1: Dual trench approach: On the left side is the shallow trench in the array and on the right the deep trench in the periphery.

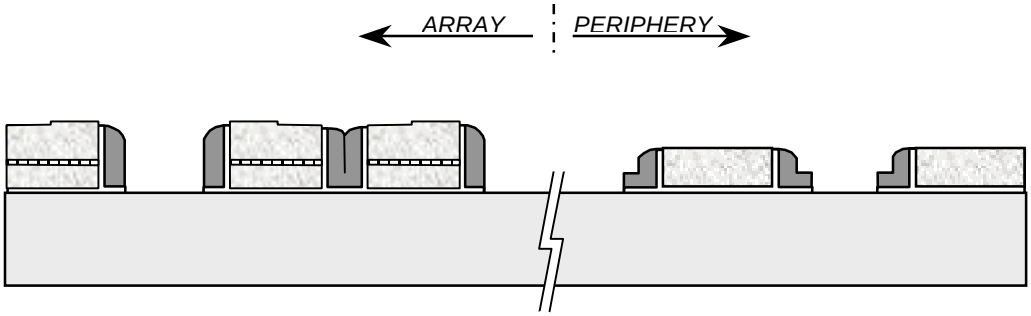

Fig. 2. Dual spacer approach: On the left is the "thin" spacer used in the flash cell and on the right is the "thick" spacer used on the high voltage transistors.

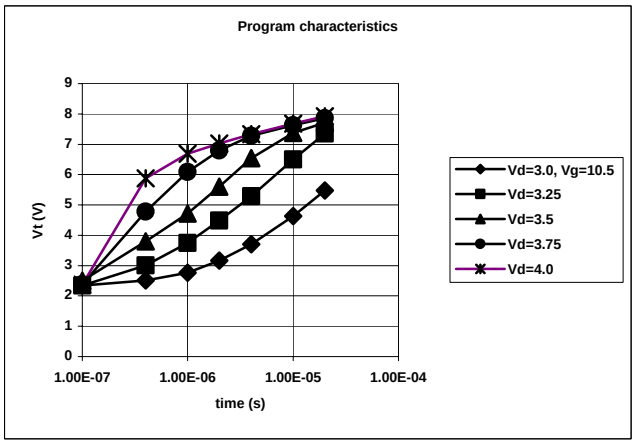

Fig.3. Programming characteristics vs drain bias show fast speed at low voltage.

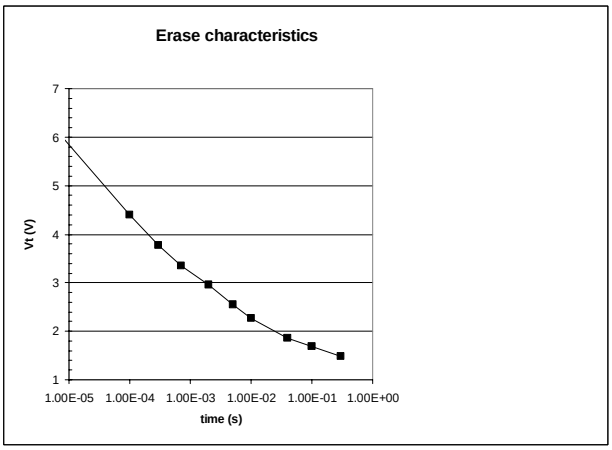

Fig.4. Erase characteristic with Vg=-8V Vwl=6V shows erase times of 100ms.

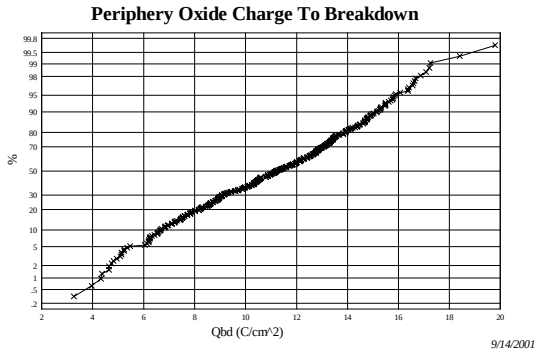

Fig.5. Periphery QBD (Charge to breakdown) shows a high quality oxide.

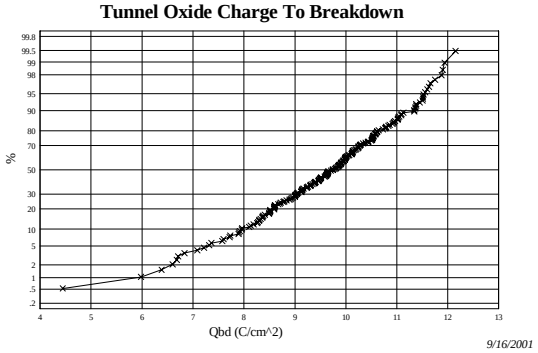

Fig.6. Tunnel oxide QBD (Charge to breakdown)shows a high quality oxide.

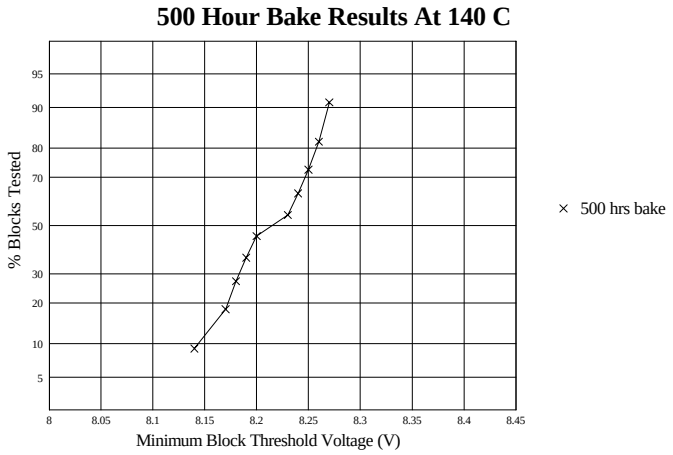

Fig.7 Program threshold voltage after 10k cycles and 500 hours bake at 140C shows very good retention

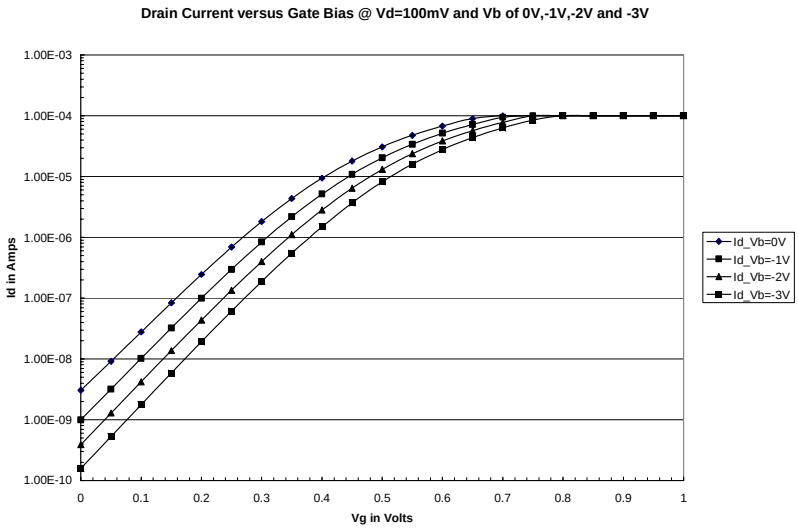

Fig.8. Periphery transistor characteristics at minimum channel length vs. back bias showing no trench corner induced kink.